Preprint typeset in JINST style - HYPER VERSION

Nonparametric Regression using the Concept of Minimum Energy

Mike Williams

Physics Department, Imperial College London, London, SW7 2AZ, United Kingdom

ABSTRACT: It has recently been shown that an unbinned distance-based statistic, the *energy*, can be used to construct an extremely powerful nonparametric multivariate two sample goodness-of-fit test. An extension to this method that makes it possible to perform nonparametric regression using multiple multivariate data sets is presented in this paper. The technique, which is based on the concept of minimizing the energy of the system, permits determination of parameters of interest without the need for parametric expressions of the parent distributions of the data sets. The application and performance of this new method is discussed in the context of some simple example analyses.

Contents

1. Introduction

The concept of *nuissance parameters*, unknown parameters whose values are of no interest but must be determined so that estimators for the parameters of interest can be obtained, is a well known one in high energy physics. The name is apt in that the presence of such parameters increases the uncertainty on the parameters of interest but has little affect on how the analysis is performed. *E.g.*, if the functional form of the probability density fuction (p.d.f.) is known, then the unkown parameters in the p.d.f. are typically determined using the least squares or maximum likelihood methods. The presence of nuissance parameters simply increases the number of parameters whose values must be determined but does not affect how the values are obtained.

The same cannot be said for *nuissance distributions*, *i.e.*, distributions whose functional form is unknown and is of no interest but (seemingly) must be determined to obtain estimators for the parameters of interest. The most common solution to this problem in high energy physics is to obtain an estimate for such a p.d.f. either by fitting a model to the data or by binning the data to obtain the integral of the p.d.f. inside the bin. Both of these methods have their limitations: the (often unknown) systematic uncertainties in the model must be propagated back to the parameters, while binning the data results in information loss that tends to increase the statistical uncertainties. Another method that is used (albeit less frequently) in high energy physics is nonparametric kernel regression (see, *e.g.*, Ref. [1]). This approach provides an unbinned data-driven way of obtaining an estimate for an unknown p.d.f.; however, its power and reliability is strongly tied to the quality of the p.d.f. estimate (which can be difficult to assess).

All of these methods share the same basic underlying idea: one must obtain an estimate of the p.d.f. to obtain estimators for the parameters of interest. In this paper I will show that this is not the case. If one has obtained multiple (possibly multivariate) data sets whose p.d.f.'s are known to be related by some set of parameters, then the values of those parameters can be determined without the need for any estimates of the p.d.f.'s themselves; all that is required is the data obtained from the p.d.f.'s. This paper is laid out as follows: the method is presented in Section 2; some example applications are given in Section 3 while a summary and discussion is provided in Section 5.

2. Method

Consider the case where n_d data sets with completely unknown p.d.f.'s, denoted by $f_1(\vec{x}) \ldots f_{n_d}(\vec{x})$, have been obtained (p.d.f.'s are normailzed such that $\int f(\vec{x})d\vec{x} = 1$). Furthermore, data has also been taken that is known to have a p.d.f. that can be written as follows:

$$f(\vec{x}) = \sum_i^{n_d} \beta_i f_i(\vec{x}), \qquad \sum_i^{n_d} \beta_i = 1, \tag{2.1}$$

where β_i are the unknown parameters of interest in the analysis. The scenario studied in this paper is that the analyst seeks to measure the parameters β_i but has no interest in the p.d.f.'s f_i.

The following test statistic correlates the difference between two p.d.f.'s at different points in a multivariate space [2, 3]:

$$\begin{aligned} T &= \frac{1}{2}\int\int \left(f(\vec{x}) - f_0(\vec{x})\right)\left(f(\vec{x}') - f_0(\vec{x}')\right)\psi(|\vec{x}-\vec{x}'|)d\vec{x}d\vec{x}' \\ &= \frac{1}{2}\int\int \left[f(\vec{x})f(\vec{x}') + f_0(\vec{x})f_0(\vec{x}') - 2f(\vec{x})f_0(\vec{x}')\right]\psi(|\vec{x}-\vec{x}'|)d\vec{x}d\vec{x}', \end{aligned} \tag{2.2}$$

where $\psi(|\vec{x}-\vec{x}'|)$ is a weighting function. T can be estimated without the need for any knowledge about the forms of f and f_0 using data sampled from the p.d.f.'s as

$$T \approx \frac{1}{n(n-1)}\sum_{i,j>i}^{n}\psi(\Delta\vec{x}_{ij}) + \frac{1}{n_0(n_0-1)}\sum_{i,j>i}^{n_0}\psi(\Delta\vec{x}_{ij}) - \frac{1}{nn_0}\sum_{i,j}^{n,n_0}\psi(\Delta\vec{x}_{ij}), \tag{2.3}$$

where $\Delta\vec{x}_{ij} = |\vec{x}_i - \vec{x}_j|$ and n (n_0) is the number of events sampled from f (f_0). In the order in which they appear in Eq. 2.3, the sums are over pairs of f events, pairs of f_0 events and pairs consisting of an f event and an f_0 event, respectively. Eq. 2.3 is simply Eq. 2.2 rewritten using the fact that $\int f(\vec{x})d\vec{x} = \int f_0(\vec{x})d\vec{x} = 1$, along with the standard Monte Carlo integration approximation.

It is straightforward to calculate T in this way once a metric is chosen that defines distance in the multivariate space (see Ref. [4] for a detailed discussion on metrics; this choice has almost no affect on the results). It is worth noting that the larger the difference is between f and f_0 the larger the expectation value of T becomes; thus, T can be used to determine the goodness of fit (g.o.f.) of the data to the hypothesis $f = f_0$ (for a more detailed discussion, see Refs. [2, 3, 4]). From this point forward I will follow Ref. [3] and refer to this test as the energy test (the name originates from the fact that if $\psi(x) = 1/x$ then Eq. 2.2 is the electrostatic energy of two charge distributions of opposite sign; the minimum energy occurs when $f = f_0$).

I will now extend the method of Ref. [2, 3] to allow for regression using multiple data sets sampled from different p.d.f.'s. Consider the following test p.d.f.:

$$f_0(\vec{x}, \vec{\beta}) = \sum_i^{n_d} \beta_i f_i(\vec{x}), \qquad \sum_i^{n_d} \beta_i = 1, \tag{2.4}$$

where β_i are unknown real parameters. The T-value that compares f and f_0 in this case is

$$\begin{aligned} T \approx \frac{1}{n(n-1)}\sum_{i,j>i}^{n}\psi(\Delta\vec{x}_{ij}) + \sum_k^{n_d}\frac{\beta_k^2}{n_k(n_k-1)}\sum_{i,j>i}^{n_k}\psi(\Delta\vec{x}_{ij}) \\ + \sum_{k,l>k}^{n_d}\frac{\beta_k\beta_l}{n_k n_l}\sum_{i,j}^{n_k,n_l}\psi(\Delta\vec{x}_{ij}) - \sum_k^{n_d}\frac{\beta_k}{nn_k}\sum_{i,j}^{n,n_k}\psi(\Delta\vec{x}_{ij}), \end{aligned} \tag{2.5}$$

where n_k is the number of events in the data set sampled from $f_k(\vec{x})$. In the order in which they appear in Eq. 2.5 the sums are over pairs of f events, pairs of f_k events, pairs consisting of an f_k event and an f_l event and pairs consisting of an f and an f_k event, respectively. The values of β_k that minimize T, $\hat{\beta}_k$, provide the best g.o.f. In the limit $n, n_k \to \infty$ Eq. 2.5 approaches Eq. 2.2 and $T = 0$ if $\hat{\beta}_k = \beta_k$ for each k. It is important to note that Eq. 2.4 is defined using properly normalized p.d.f.'s. If unnormalized functions are used, then the estimators are related to the true values by $\hat{\beta}_k = \beta_k \int f_k(\vec{x}) d\vec{x}$. The uncertainty on the fit parameter values is easily obtained using a resampling technique (*e.g.*, bootstrapping).

One might think that calculating T takes so much CPU time that running this fit is not practical; however, a careful inspection of Eq. 2.5 reveals that the CPU-intensive components of T, the $\sum \psi(\Delta\vec{x})$ terms, do not depend on the β_k values. Thus, these terms only need to be calculated once (which can be done prior to running the fit). A limitation of this method, however, is that if a p.d.f. contributes negative probability to f_0, then the fits may be unstable due to fact that it is not possible to constrain f_0 to be non-negative for all $\vec{x}$. In many cases it will be easy to tell if this is a problem or not. Otherwise, I recommend performing a Monte Carlo study prior to using this method in such situations. *N.b.*, in Section 3.2 an example is presented where $n_d = 3$ and one p.d.f. contributes negative probability to f_0 and the method still works properly (in that example, the known relationships between the p.d.f.'s makes it easy to determine that the p.d.f. is non-negative everywhere even though one term contributes negative probability).

In the next section I will present two examples to help illustrate how the method works. The main goal will be to not only demonstrate how to use the method but to also solidify in the reader's mind what the $\hat{\beta}_k$ values represent. I conclude this section by restating the main result of this work: estimators for the parameters β_k in Eq. 2.1 can be obtained by minimizing Eq. 2.5 using data sets sampled from f_k without the need for any knowledge about the p.d.f.'s themselves.

3. Example Applications

The examples I will present in this section are meant to demonstrate how the method works. From a physics perspective, these are not the most interesting applications of the method; however, they are effective at illustrating how to use the method and how to interpret the meaning of the fit parameters. I will first present a very simple univariate example followed by an example using Dalitz plots. The weighting function that will be used for the examples below is $\psi(x) = -\log(x+\varepsilon)$, where ε is of the order of the inverse of the total number of events in all samples combined (ε simply guards against an infinite contribution due to two extremely close events; the exact value is not important). See Section 5 for discussion on the choice of weighting function.

3.1 Univariate Example

Consider the case where one has three data sets sampled from the p.d.f.'s $f_c(x) = 1$, $f_l(x) = 2x$ and $f(x) = (ax+b)/(a/2+b)$ on the interval $x \in [0,1)$. The sample sizes of these data sets are denoted by n_c, n_l and n, respectively. The p.d.f. f can be rewritten in terms of the other two p.d.f.'s as

$$f(x) = \frac{\beta_1 f_l(x) + \beta_0 f_c(x)}{\beta_1 + \beta_0}, \tag{3.1}$$

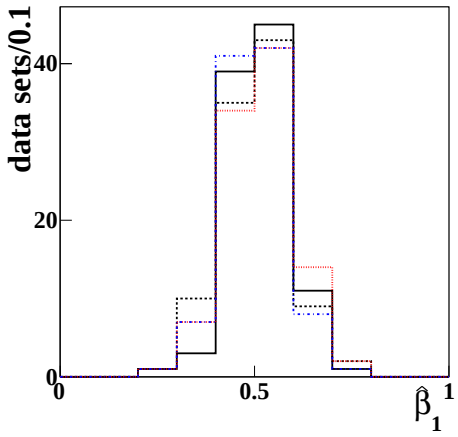

Figure 1. $\hat{\beta}_1$ values obtained for the ensemble of data sets described in Section 3.1 using the energy test for $n_{c,l} = 1$k (black, dashed) and $n_{c,l} = 10$k (black, solid). Results obtained using the binned χ^2 test (red, dotted) and maximum likelihood test (blue, dash-dotted) are also shown.

where $\beta_1 = a/2$, $\beta_0 = b$ and the normalization requirement on the β values has been made explicit. The values $\beta_1/(\beta_1+\beta_0)$ and $\beta_0/(\beta_1+\beta_0)$ represent the fraction of f's probability associated with f_l and f_c, respectively. *E.g.*, if f_l and f_c represent signal and background p.d.f.'s then the signal purity is given by $\beta_1/(\beta_1+\beta_0)$. The factor of 2 in β_1 arises because the function x is not a properly normalized p.d.f. on this interval.

The fit procedure is simple: first, all of the $\sum \psi(\Delta x_{ij})$ terms in Eq. 2.5 need to be calculated. This can be time consuming for large data sets but it only needs to be done once. The test p.d.f. is then written as

$$f_0(x,\vec{\beta}) = \frac{\beta_1 f_l(x) + \beta_0 f_c(x)}{\beta_1+\beta_0}, \tag{3.2}$$

where no knowledge about the forms of f_l and f_c is assumed to be known and the normalization requirement on the β values has again been made explicit. Calculating T using Eq. 2.5 is then straightforward for any given values of β_1 and β_0. The estimators for β_i are the values that minimize T, $\hat{\beta}_i$.

I generated an ensemble of 100 data sets from Eq. 3.1 using the values $\beta_0 = \beta_1 = 1/2$ (or, equivalently, $a = 1, b = 1/2$). I chose $n = 500$ and two different values for n_c and n_l: $n_{c,l} = 1$k and $n_{c,l} = 10$k. Figure 1 shows the $\hat{\beta}_1$ values ($\hat{\beta}_0$ is just $1-\hat{\beta}_1$) obtained for these data sets using the minimum energy, χ^2 and maximum likelihood tests (statistics are given in Tab. 1). The χ^2 and maximum likelihood tests were performed by fitting the function $f_0(x) \propto \beta_1 x + \beta_0$ to the data sets sampled from f; *i.e.*, the data sets sampled from f_c and f_l were not used and, instead, the functional form of f was assumed to be known. Even though I *cheated* (*i.e.*, used information I assumed I do not actually have access to) using these tests, the performance of the energy test is still comparable. The method works: I have determined the fraction of probability in f from f_c and f_l without having any knowledge of the forms of f_c or f_l and without trying to approximate f_c and f_l using the data (*e.g.*, using a kernel-based method).

As stated above, the uncertainty on the fit parameters must be obtained using a data resampling technique. There are many such techniques available; I chose to use bootstrapping. This technique

χ^2 test		$\log\mathscr{L}$ test		energy test			
				$n_{c,l} = 1$k		$n_{c,l} = 10$k	
μ	r.m.s.	μ	r.m.s.	μ	r.m.s.	μ	r.m.s.
0.52	0.09	0.50	0.08	0.50	0.09	0.51	0.08

Table 1. $\hat{\beta}_1$ statistics obtained for the ensemble of data sets described in Section 3.1 using the χ^2, maximum likelihood and energy tests.The data sets were sampled from a p.d.f. with a value of 0.5.

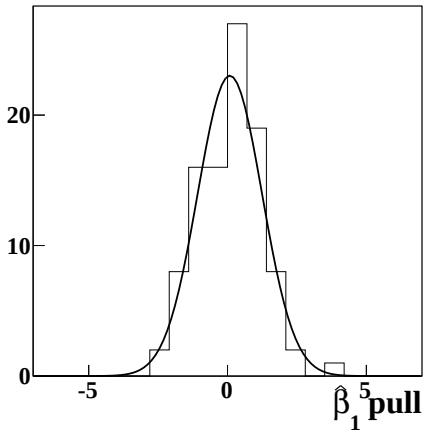

Figure 2. $\hat{\beta}_1$ pull distribution obtained for the $n_{c,l} = 10$k data sets where the parameter uncertainties are estimated using the bootstrapping technique. The solid line shows a fit to a gaussian distribution; the fit yields $\mu = 0.08 \pm 0.13$ and $\sigma = 1.16 \pm 0.11$.

involves making $n_{\rm boot}$ (I chose 100) resampled data sets from each f, f_c and f_l data set. These *bootstrap copy* data sets are produced by sampling with replacement (thus, the sample sizes are unchanged) from the originals. The fit parameter $\hat{\beta}_1$ is then computed for each of the $n_{\rm boot}$ bootstrap data sets; the standard deviation of this distribution is used as an estimate for the uncertainty on $\hat{\beta}_1$. Figure 2 shows the $\hat{\beta}_1$ pull distribution obtained for the ensemble of data sets described in the previous paragraph. The uncertainty on $\hat{\beta}_1$ was determined for each data set in the ensemble using bootstrapping. The pull distribution is consistent with the expected standard normal distribution; thus, I conclude that the bootstrap method produces reliable estimates for the uncertainty on the $\hat{\beta}_k$ values. For more information on bootstrapping, see Ref. [5].

3.2 Multivariate Example

I will now consider a multivariate example involving Dalitz plots. The decay $D \to K_S\pi^+\pi^-$ is allowed for $D = D^0$, $\bar{D}^0$ and any superposition of D^0 and $\bar{D}^0$. Experimentally, flavor-tagged samples (*i.e.*, pure D^0 or $\bar{D}^0$) can be obtained using the decays $D^{*+} \to D^0\pi^+$ and $D^{*-} \to \bar{D}^0\pi^-$ since the charge of the slow pion tags the flavor of the D. Such data sets with sample sizes of $\mathscr{O}(100\text{k})$ have been obtained at Belle [6]. *CP* eigenstates can be obtained using quantum correlated $D^0\bar{D}^0$ pairs produced in the reaction $e^+e^- \to D^0\bar{D}^0$. *CP*-tagged data sets have been obtained at CLEO with sample sizes of several hundreds of events [7], while sample sizes of $\mathscr{O}(1000)$ have been recorded (but not yet published) at BES [8].

I will proceed under the assumption of no *CP* violation in the neutral *D* system, which is known to be valid to a high level of precision. I will denote the quantum mechanical amplitudes for the decays $D^0, \bar{D}^0 \to K_S\pi^+\pi^-$ as follows:

$$\mathscr{A}_{\bar{D}^0\to K_S\pi^+\pi^-}(\vec{x}) \equiv \bar{\mathscr{A}}(\vec{x}), \tag{3.3}$$

$$\mathscr{A}_{D^0\to K_S\pi^+\pi^-}(\vec{x}) \equiv \mathscr{A}(\vec{x}), \tag{3.4}$$

where $\vec{x} = (m_+^2, m_-^2)$ and m_+^2 and m_-^2 are the invariant masses of the $K_S\pi^+$ and $K_S\pi^-$ systems, respectively. The flavor-tagged p.d.f.'s are then given by

$$\bar{f}(\vec{x}) = |\bar{\mathscr{A}}(\vec{x})|^2/\mathscr{I}, \tag{3.5}$$

$$f(\vec{x}) = |\mathscr{A}(\vec{x})|^2/\mathscr{I}, \tag{3.6}$$

where $\mathscr{I} = \int |\bar{\mathscr{A}}(\vec{x})|^2 d\vec{x} = \int |\mathscr{A}(\vec{x})|^2 d\vec{x}$ (which is valid in the absence of *CP* violation). The amplitudes for the case where the *D* is tagged to be in a *CP* eigenstate are given by

$$\mathscr{A}_\pm(\vec{x}) = \frac{1}{\sqrt{2}}[\bar{\mathscr{A}}(\vec{x}) \pm \mathscr{A}(\vec{x})]. \tag{3.7}$$

The *CP*-tagged p.d.f.'s are thus

$$f_\pm(\vec{x}) = \left(|\bar{\mathscr{A}}(\vec{x})|^2 + |\mathscr{A}(\vec{x})|^2 \pm 2|\bar{\mathscr{A}}(\vec{x})||\mathscr{A}(\vec{x})|\cos\Delta\theta(\vec{x})\right)/\mathscr{I}_\pm, \tag{3.8}$$

where

$$\mathscr{I}_\pm = 2\left(\mathscr{I} \pm \int |\bar{\mathscr{A}}(\vec{x})||\mathscr{A}(\vec{x})|\cos\Delta\theta(\vec{x})d\vec{x}\right) \tag{3.9}$$

and $\Delta\theta(\vec{x})$ is the phase difference between $\bar{\mathscr{A}}(\vec{x})$ and $\mathscr{A}(\vec{x})$ at each point $\vec{x}$. I generated an ensemble of 100 data sets for each flavor- and *CP*-tagged decay using the model of Ref. [6] (the amplitudes were evaluated using the `qft++` package [9]). Figure 3 shows an example of a data set of each tagging type. The sample sizes were chosen to be $n = \bar{n} = 100$k and $n_\pm = 1000$ for the D^0, $\bar{D}^0$ and $D_{CP\pm}$ decays, respectively.

As an example of using the method presented in this paper, consider the case where one has collected both flavor-tagged and *CP*-tagged data sets. I will assume that the only knowledge about $f(\vec{x})$ and $\bar{f}(\vec{x})$ is the fact that $\int |\bar{\mathscr{A}}(\vec{x})|^2 d\vec{x} = \int |\mathscr{A}(\vec{x})|^2 d\vec{x}$; the functional forms of f and $\bar{f}$ are completely unknown. I will also assume that the *CP*-odd tagged data set is known to follow Eq. 3.8; however, only the relative coefficients of $|\bar{\mathscr{A}}(\vec{x})|^2$, $|\mathscr{A}(\vec{x})|^2$ and $|\bar{\mathscr{A}}(\vec{x})||\mathscr{A}(\vec{x})|\cos\Delta\theta(\vec{x})$ are known (nothing is known about their functional forms). For the *CP*-even data set, I assume that its p.d.f. can be written as

$$f_+(\vec{x},\vec{\alpha}) \propto \alpha_0|\bar{\mathscr{A}}(\vec{x})|^2 + \alpha_1|\mathscr{A}(\vec{x})|^2 + \alpha_2|\bar{\mathscr{A}}(\vec{x})||\mathscr{A}(\vec{x})|\cos\Delta\theta(\vec{x}). \tag{3.10}$$

The goal of the analysis is to test the quality of the *CP* tagging by determining the α_i values (up to an arbitrary normalization factor) using only the available knowledge about the data collected, which does not include any knowledge about the functional forms of $\bar{\mathscr{A}}(\vec{x})$ or $\mathscr{A}(\vec{x})$. A deviation from the *CP*-even values, $\vec{\alpha} \propto (1,1,2)$, would signify a problem in the *CP* tagging.

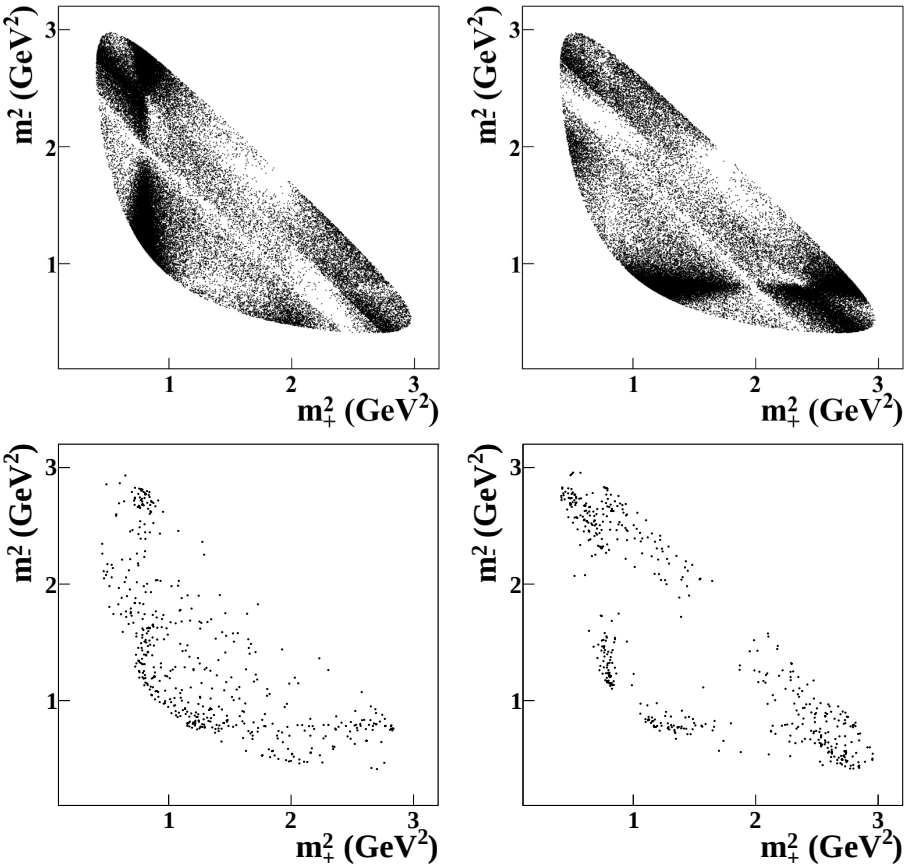

Figure 3. Dalitz plots for $D \to K_S\pi^+\pi^-$ for (top left) $\bar{D}^0$, (top right) D^0, (bottom left) D_{CP+} and (bottom right) D_{CP-}.

The energy test can be used to determine the β_i values in the following test p.d.f. using the same procedure described in the previous section:

$$f_0(\vec{x},\vec{\beta}) = \beta_0\bar{f}(\vec{x}) + \beta_1 f(\vec{x}) + \beta_2 f_-(\vec{x}), \qquad \sum\beta_i = 1, \tag{3.11}$$

where f, $\bar{f}$ and f_- are the flavor-tagged and CP-odd-tagged p.d.f.'s, respectively (as defined above). Comparing Eq. 3.11 and Eq. 3.10 one can see that the $\vec{\beta}$ values are related to the $\vec{\alpha}$ values by

$$\vec{\alpha} \propto (\beta_0 + \beta_2\mathscr{I}/\mathscr{I}_-, \beta_1 + \beta_2\mathscr{I}/\mathscr{I}_-, -2\beta_2\mathscr{I}/\mathscr{I}_-)\,. \tag{3.12}$$

Thus, to determine the coefficients of $|\mathscr{A}\bar{}(\vec{x})|^2$, $|\mathscr{A}(\vec{x})|^2$ and $|\mathscr{A}\bar{}(\vec{x})||\mathscr{A}(\vec{x})|\cos\Delta\theta(\vec{x})$ for the CP-even data set (*i.e.*, to determine $f_+(\vec{x})$ using the energy test) requires the value of $\mathscr{I}_-/\mathscr{I}$ to be known (or obtainable). At first glance one might think that this is a showstopper; however, the ratio of these integrals can be obtained from the sample sizes obtained for the various data sets.

Ref. [7] measured both the flavor- and CP-tagged yields using quantum-correlated $D^0\bar{D}^0$ pairs; thus, $\mathscr{I}_-/\mathscr{I}$ can be obtained using

$$\frac{\mathscr{I}}{\mathscr{I}_-} = \frac{(n'/n_D + \bar{n}'/n_{\bar{D}})/2}{2n_-/n_{CP-}}, \tag{3.13}$$

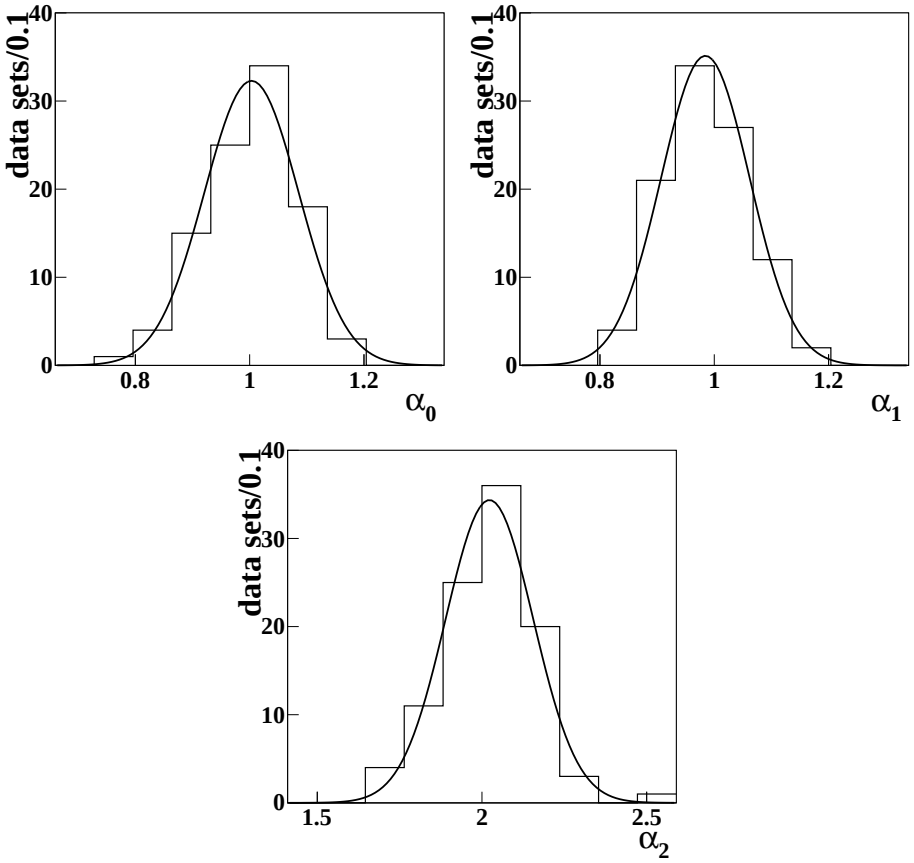

Figure 4. Values obtained for $\vec{\alpha}$ in Eq. 3.10 using the energy test. The solid lines shows fits of gaussian distributions to the data. The resulting means and widths are: $\mu_0 = 1.00 \pm 0.01$, $\sigma_0 = 0.07 \pm 0.01$; $\mu_1 = 0.98 \pm 0.01$, $\sigma_1 = 0.07 \pm 0.01$ and $\mu_2 = 2.02 \pm 0.02$, $\sigma_2 = 0.14 \pm 0.02$.

where n_D, $n_{\bar{D}}$ and n_{CP-} are the total number of D^0, $\bar{D}^0$ and *CP*-odd tagged quantum-correlated $D^0\bar{D}^0$ pairs, respectively, and $n', \bar{n}'$ are the number of flavor-tagged $D^0, \bar{D}^0 \to K_S\pi\pi$ decays observed at CLEO (not the ones observed at Belle; the Belle data are used in the fits because they have larger sample sizes). The sample sizes are $n_D \sim 100n'$, $n_{\bar{D}} \sim 100\bar{n}'$ and $n_{CP-} \sim 100n_-$. The uncertainties on the various yield ratios in Eq. 3.13 are binomial; thus, the statistical uncertainty on $\mathscr{I}/\mathscr{I}_-$ is negligible due to the fact that the total number of *CP*-tagged quantum-correlated $D^0\bar{D}^0$ pairs is $100\times$ larger than the *CP*-tagged $D \to K_S\pi\pi$ samples.

Figure 4 shows the results obtained using the energy test to determine $\vec{\beta}$ and Eq. 3.12 to convert these into $\vec{\alpha}$ values. The results are in excellent agreement with the true values $\vec{\alpha} = (1,1,2)$. The relative statistical uncertainty obtained on the α values is 7% (see Section 5 for discussion on improving this resolution). Looking at Fig. 3 one can imagine that obtaining a model p.d.f. for this analysis with a small systematic uncertainty requires a lot of work. Since the goal of this example was to simply obtain estimators for $\vec{\alpha}$ (to test the quality of the *CP* tagging), there is no reason to build such a model. Instead, one can just use the energy test.

4. Background & Efficiency

In these examples I have ignored two very important experimental issues: (1) the existence of background events and (2) the presence of non-uniform detector efficiency effects. Both of these can be accounted for in this method by introducing a weighting factor, w, for each event. The weight factor is simply given by $w_i = P_S^i / P_D^i$, where P_S^i and P_D^i are the probabilities that the event is signal (and not background) and is detected, respectively. Eq. 2.5 would then need to be rewritten as follows:

$$T \approx \frac{1}{W^2} \sum_{i,j>i}^{n} w_i w_j \psi(\Delta\vec{x}_{ij}) + \sum_{k}^{n_d} \frac{\beta_k^2}{W_k^2} \sum_{i,j>i}^{n_k} w_i w_j \psi(\Delta\vec{x}_{ij}) + \sum_{k,l>k}^{n_d} \frac{\beta_k \beta_l}{W_k W_l} \sum_{i,j}^{n_k,n_l} w_i w_j \psi(\Delta\vec{x}_{ij}) - \sum_{k}^{n_d} \frac{\beta_k}{W W_k} \sum_{i,j}^{n,n_k} w_i w_j \psi(\Delta\vec{x}_{ij}), \quad (4.1)$$

where W and W_k are the sum of the weight factors for the f and f_k data sets, respectively. If $w_i = 1$ for all events in all data sets then Eq. 2.5 is recovered (up to a negligible change, $n(n-1) \to n^2$, in the same data set sums; this change has been made solely for ease of notation and could easily be omitted if desired). *N.b.*, in many cases these weights can be left out. It is up to the analyst to decide when these factors are important and to include them when necessary.

5. Summary & Discussion

In this paper I have shown that the concept of minimum energy can be used to carry out nonparametric regression. The parameter values that are obtained give the fractional probability associated with each p.d.f. For many analyses, *e.g.*, signal-background subtraction, this is sufficient. To convert these to coefficients of unnormalized functions requires inputting the ratio of the integrals of these functions. These ratios can often times be obtained from measured event yields (an example of which was given in Section 3.2). I have also shown that data resampling methods, *e.g.*, bootstrapping, can be used to obtain good estimates of the uncertainties on the parameter values.

It is well known that in many binned analyses specialized binning schemes can be developed to enhance the resolution on the parameters of interest of the χ^2 test. The same is true for the energy test with regards to the weighting function $\psi(x)$. In this paper I chose a simple generic $\psi(x)$ because the goal was to demonstrate how the method works; however, for many analyses it will be possible to obtain a better resolution by choosing a $\psi(x)$ tailored to the problem being analyzed. Refs. [2, 3] discuss several possible choices for $\psi(x)$ and there are undoubtedly many other possibilities as well; however, for most analyses (where the p.d.f.'s vary in a smooth and slow way) the choice used in this paper is likely to be close to optimal (whatever that may be).

Acknowledgements

I would like to thank Ulrik Egede, Tim Gershon and Vladimir Gligorov for discussions. This work is supported by the Science and Technology Facilities Council (United Kingdom) under grant number ST/H000992/1.

References

[1] W. Härdle, *Applied nonparametric regression*, Cambridge University Press, New York (1990).

[2] L. Baringhaus and C. Franz, *On a new multivariate two-sample test,* J. Multivariate Anal. **88** (2004) 190-206.

[3] B. Aslan and G. Zech, *New test for the multivariate two-sample problem based on the concept of minimum energy,* Stat. Comp. Simul. **75**, Issue 2 (2004) 109-119; B. Aslan and G. Zech, *Statistical energy as a tool for binning-free, multivariate goodness-of-fit tests, two-sample comparison and unfolding,* Nucl. Instrum. Methods **A537** (2005) 626-636.

[4] M. Williams, *How good are your fits? Unbinned multivariate goodness-of-fit tests in high energy physics*, JINST **5**, (2010) P09004.

[5] B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapham and Hall, London (1993).

[6] A. Poluektov *et al.* [Belle Collaboration], *Measurement of ϕ_3 with Dalitz plot analysis of $B^{(*)} \to D^{(*)}K^{\pm}$ decay*, Phys. Rev. D **70**, (2004) 072003.

[7] R.A. Briere *et al.* [CLEO Collaboration], *First model-independent determination of the relative strong phase between D^0 and $\bar{D}^0 \to K_S\pi^+\pi^-$ and its impact on the CKM angle γ/ϕ_3 measurement*, Phys. Rev. D **80**, (2009) 032002.

[8] R.A. Briere, private communication.

[9] M. Williams, *Numerical object oriented quantum field theory calculations,* Comp. Phys. Comm. **180**, (2009) 1847.